NASA Contractor Report 198238

ICASE Report No. 95-78

ICASE

ARCHITECTURE AND PERFORMANCE ANALYSIS OF DIRSMIN: A FAULT-TOLERANT SWITCH USING DILATED REDUCED-STAGE MIN

Arun K. Somani
Tianming Zhang

NASA Contract No. NAS1-19480
November 1995

Institute for Computer Applications in Science and Engineering
NASA Langley Research Center
Hampton, VA 23681-0001

Operated by Universities Space Research Association

National Aeronautics and Space Administration

Langley Research Center
Hampton, Virginia 23681-0001

Architecture and Performance Analysis of DIRSMIN: A Fault-Tolerant Switch using Dilated Reduced-Stage MIN

*Arun K. Somani** *and Tianming Zhang*
Department of Electrical Engineering and
Department of Computer Science and Engineering
University of Washington
Seattle, WA 98195

Abstract

We develop and analyze a dilated high performance fault tolerant fast packet multistage interconnection network (MIN) in this paper. In this new design, the links at the input and the output stages of a dilated banyan-based MIN are rearranged to create multiple routes for each source-destination pair in the network after removing one stage in the network. These multiple paths are link- and node-disjoint. Fault tolerance at low latency is achieved by sending multiple copies of each input packet simultaneously using different routes and different priorities. This guarantees that high throughput is maintained even in the presence of faults. Throughput is analyzed using simulation and analysis and we show that the new design has considerably higher performance in the presence of a faulty switching element (SE) or link in comparison to dilated networks. We also analyze the reliability and show that the new design has superior reliability in comparison to competing proposals.

*An earlier version of this paper containing reliability analysis appeared in Proc. of IEEE INFOCOM, 1995. The research reported in this paper was supported in part by the NSF under Grant NCR 9103485. This work was also in part supported by NASA under NAS1-19480 while the author was on sabbatical at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA 23681

1 Introduction

High dependability is required in communication for multiprocessor and communication systems. For example, high bandwidth transmission systems to carry high volumes of video, voice, and data in Broadband Integrated Services Digital Networks (B-ISDN) using Asynchronous Transfer Mode (ATM) are becoming common. It remains a challenge to integrate reliability while maintaining high throughput in switches for B-ISDN. Most switching architectures proposed [1]-[9] use self-routing, space-switching, and internal nonblocking paradigms. Banyan-based Multistage Interconnection Networks (MINs) such as Omega and Generalized-cube (GC) [1] have received considerable attention due to their favorable cost/performance ratio.

High blocking probability and low throughput, however, greatly limit their capability of handling fast packet switching due to internal contention. Many schemes to reduce blocking probability and increase throughput have been developed [2]-[9]. Batcher-banyan networks [2], internal speed up [3], replication in series (Tandem banyan) [4, 5] or in parallel [6], link dilation in MIN [7, 8], and multi-priority traffic [9] are some of them. The maximum throughput achievable with head-of-line (HOL) collision [10] remains at 0.58. None of these can tolerate failures. To overcome this problem, redundant paths in a MIN are provided by adding extra stages or links [11]-[16]. The performance of these networks in the presence of faults is affected adversely and is too low for fast packet switching for B-ISDN. Altogether, most of these methods are not suitable for fast-packet switching.

Our proposal to achieve fault tolerance without sacrificing performance is to use dilated networks and rearrange input/output links to provide redundant paths between a source/destination pair. We show that by modifying the input and output stages of a dilated network, high performance, low cost, and fault-tolerance can be achieved at the same time. In particular, we discuss the role of dilation in fault tolerance in Section 2 and develop a space-division fast packet design, called the dilated reduced-stage Multistage Interconnection Network (DIRSMIN) in Section 3. We evaluate the performance of this network and present our simulation results in Section 4. In Section 5, we establish the analytical model to compute the performance of DIRSMIN. In

Section 6, we develop the reliability model of DIRSMIN. In Section 7, we summarize the main results and discuss some possible extensions.

2 Dilation and Fault Tolerance

An $N \times N$ multistage banyan network consists of $\log N$ (base 2) stages as shown in Figure 1a. Each stage has $N/2$ 2×2 SEs. Different interconnection functions can be used to yield different topologies such as generalized cube and Omega network as shown in Figures 1a and 1b, respectively. An Omega network can be redrawn, as shown in Figure 1e, to show that it is equivalent to a cube network. So we consider only Omega networks in our discussion.

To reduce contention and improve the performance of each link, an Omega network can be dilated by d to form a d-dilated network [7, 8]. It has been shown by us and other researchers that to meet the low-latency and high performance requirement, $d = \log \log N$ dilation is optimal. All stages use d-dilated 2×2 or $2d \times 2d$ SEs. A packet entering a SE may exit using any of the d links going to the desired SE in the next stage. The cost considerations limit the degree of dilation. Kumar and Jump [8] have shown that the dilated networks always have higher performance than other comparable schemes. Dilation by itself does not provide the fault tolerance in the network in the presence of faulty SEs. The dependability can be provided in a dilated banyan network by modifying and rearranging the input and output connections to create multiple paths using the strategy described below.

To tolerate input/output link failures, each source and destination must be connected to multiple ports. Each source may feed data through up to p different input ports and each destination receives data from q different output ports. In a non-dilated banyan network, we need an in-mux $N \times p$ to N stage and an out-demux N to $N \times q$ stage to match the number of input and output ports of the network with the number of source/destinations as shown in Figure 1c. The banyan-based MIN remains unchanged. This structure employing p input and q output links is called an extra-link MIN or ELMIN(p, q).

By suitably choosing the I/O connections, we can create up to $p \times q < N/2$ possible paths between each source/destination pair, at least $min(p, q)$ of them entirely independent. Let bit patterns $s_{n-1} \cdots s_0$ and $d_{n-1} \cdots d_0$ be the binary strings representing the addresses of source S and destination D. Typically, in MINs, source $S = s_{n-1} \cdots s_0$ is connected to network input $I = s_{n-1} \cdots s_0$. Similarly, a destination $D = d_{n-1} \cdots d_0$ is connected to network output $O = d_{n-1} \cdots d_0$. In ELMIN(p, q), a source S is connected to p (assuming p is a power of two) input ports of the network whose addresses are derived by using all possible combinations of the least significant $\log p$ bits of the source address. Similarly, a destination receives from q output ports whose addresses are derived by using all possible combinations of the most significant $\log q$ bits of its address. An ELMIN$(2, 2)$ derived using an Omega network is shown in Figure 1d without in-mux/out-demux stages. The in-mux and out-demux can be removed by substituting $2p \times 2$ and $2 \times 2q$ SEs in the input and the output stages, respectively. The most appropriate values seem to be $p = q = 2$.

Multiple path MINs can also be used to improve the performance of the switch by balancing the load over these paths. In this scheme, a cell is sent through all available paths with different priorities. In case of contention, a cell being routed on the primary path has higher priority. Thus, the basic capabilities of a MIN are not affected. However, additional cells may be routed through the secondary paths, improving the overall performance. We use this scheme in dilated banyan networks after rearranging the links to achieve higher throughput in our design, presented in the next section.

3 Dilated Reduced-Stage MIN (DIRSMIN)

To tolerate failure of an internal SE node, a combination of dilation and the ELMIN type connection scheme offers a very attractive option [17]. Because of dilation, extra links coming from source and to destination nodes can be directly connected to SEs without incorporating any in-mux or out-mux stages as shown in Figure 1d. Also, due to multiple ports connecting to a destination, the switching of the most significant bit in routing in the banyan network in the

first stage is not required. Thus, the first stage can be eliminated. This $\log N - 1$ stage dilated MIN is called a Dilated Reduced Stage MIN (DIRSMIN) network. The stages from the input to the output are numbered as stage $(n-2)$ to stage 0. The last stage uses dilated 2×4 SEs to provide adequate routing. The rest of the stages use dilated 2×2 SEs. DIRSMIN reduces the delay in the network as there is one less switching stage. Moreover, interestingly in our simulations we noticed that the throughput improves due to reduction in stages.

3.1 Input/Output Connections.

We develop an example design with four inputs and two dilated output links to provide adequate performance and fault tolerance. The links in non-dilated banyan are numbered from 0 to N-1 from top to bottom at the input and output of each stage independently. In ELMIN, each source is connected to the two network ports specified by $s_{n-1}s_{n-2}\cdots s_1 s_0$ and $s_{n-1}s_{n-2}\cdots s_1\overline{s_0}$. After the shuffle, these ports are connected to links $s_{n-2}\cdots s_1 s_0 s_{n-1}$ and $s_{n-2}\cdots s_1\overline{s_0}s_{n-1}$, respectively. Each of the corresponding SEs can switch this link to two output links that are specified by $s_{n-2}\cdots s_1 s_0 s_{n-1}$ and $s_{n-2}\cdots s_1 s_0\overline{s_{n-1}}$ for the original connection and $s_{n-2}\cdots s_1\overline{s_0}s_{n-1}$, and $s_{n-2}\cdots s_1\overline{s_0 s_{n-1}}$ for the second connection in ELMIN. The links go through a shuffle at the output of the first stage. In DIRSMIN, we connect the four output links from each source directly to the corresponding positions in the new first stage (second stage earlier) with different priorities as shown in Figure 1f.

The network can be partitioned into four subnetworks from Stage $n-2$ to Stage 1 as shown in Figure 1g. The SEs in stage 0 combine outputs from two subnetworks and then route to the destination. This partitioning is shown in Figure 1g. The subnetworks are numbered from 0 to 3. Notice the order of numbering. Let the subnetwork number be denoted by $n_1 n_0$. Each source is connected to exactly one input in each subnetwork. If the links in each subnetwork are renumbered from 0 to $N/4 - 1$, then a source $S = s_{n-1}s_{n-2}\cdots s_1 s_0$ is connected to link $s_{n-3}\cdots s_1 s_{n-2}$ in each subnetwork. A destination $d_{n-1}\cdots d_0$ is connected to a pair of conjugate SEs, $d_{n-1}d_{n-2}\cdots d_1$ and $\overline{d_{n-1}}d_{n-2}\cdots d_1$.

In each subnetwork, there are exactly N packets, divided equally among four priorities. In one time slot, each destination can receive up to $2d$ packets from the two conjugate SEs. The number of output links of a SE at the last stage of the modified DIRSMIN does not have to be the same as the dilation degree and can be a design parameter d' decided based on the desired performance. We use the notation (d, d')-DIRSMIN to indicate a DIRSMIN with d dilation from stages $(n-2)$ to 0, and d' dilation for the four output ports of each 2×4 SE in the last stage. Thus, $2d'$ is the total number of links connected to each destination and d' can be varied in range of $1 \leq d' \leq d$. Notice that a $(d, d/2)$-DIRSMIN has the same number of links connected to each destination as a d-dilated Omega network but has one less stage.

3.2 Operation and Fault Tolerance.

The network is operated in a time synchronized fashion as required in the ATM standard. In each time slot, every source sends four copies of an input packet with four different priorities to the four links with the destination as the routing tag. The priority of a packet from that source in subnetwork $n_1 n_0$ is given by $s_0 s_{n-1} \oplus n_1 n_0$. For example, for the case of 16×16, nodes 0, 1, 8, and 9 are all connected to dilated link 0 in each subnetwork. However, the priority of source 0 is 0, 1, 2, and 3, and the priority of source 9 is 3, 2, 1, and 0 in subnetworks 0, 1, 2, and 3, respectively. We refer to priority 0 link as the primary link, priority 1 link as the secondary link, and so on. Priority 0 is the highest priority. The routing of packets is governed by the destination address bits d_{n-2} to d_1 in stages $(n-2)$ to 1, respectively. SEs in the last stage route packets to 4 different output ports using bits $d_0 d_{n-1}$. Contentions can occur within each SE when the number of packets destined to one particular output port of a SE exceeds the dilation degree of that port. In such a situation, packets are dropped following the priority order. In case of contention among the same priority packets, packets are dropped randomly. If more than one copy of the same packet arrive at the last stage, only the copy with the highest priority is transmitted to the destination.

The redundancy graph of the four priority copies from a source to a destination is shown in

Figure 1h. At the last stage, priority 0 and priority 2 copies of a packet from the same input reach the same SE. Similarly, priority 1 and priority 3 copies also reach another SE. Low priority copies will be lost if there is contention in the last stage. Thus, the probability of priority 2 and priority 3 packets that pass through the network and become the only successful copies is small. This is also seen in our simulation and analytical results. Accordingly, a two priority scheme may be used, where each source sends out only two copies in each time slot if that level of fault tolerance is acceptable. In this case, each incoming packet has two routes instead of four and the scheme may be less robust than the four priority scheme under multiple faults. By sending redundant copies to four SE-disjoint subnetworks in one time slot, DIRSMIN can tolerate at least 3 SE faults in stages $n-2$ to 1 and at least 1 SE fault at the last stage. It is robust in the presence of more SE faults in the network. In particular, the network works well even when one whole subnetwork becomes faulty.

3.3 Implementation Issues

A DIRSMIN is designed based on a d-dilated banyan network. Each SE from stage $(n-2)$ to 1 in a (d, d')-DIRSMIN has the same size, $2d \times 2d$, as a SE in a d-dilated banyan network. In a (d, d)-DIRSMIN, the size of the SEs at the last stage are $2d \times 4d$, which is twice as big as the size of the SEs used in a d-dilated banyan network. But a (d, d)-DIRSMIN has one less stage than a d-dilated banyan network. One way to compare the costs of different networks is to estimate SE complexity, defined by the total number of cross points in the switch. The number of cross points in a (d, d)-DIRSMIN is $(\log N - 2) \times N/2 \times 2d \times 2d + N/2 \times 2d \times 4d = 2Nd^2 \log N$. This is the same as in a d-dilated Omega network. Thus, the two networks are equivalent in terms of switch complexity. A (d, d')-DIRSMIN with $d' < d$ has fewer crosspoints than the corresponding Omega network with dilation d.

4 Performance Using Simulation Techniques

We compare the performance of DIRSMIN and dilated omega network using simulation following commonly used assumptions. As stated earlier dilated Omega network has been shown to yield the best performance/cost among the competing architectures. It has also been shown that ELMIN type networks perform much better under non-uniform traffic conditions in [15]. So we will not deal with that here. In the simulation, we use a Bernoulli process with parameter λ to describe the arrival of packets at a source node and assume that the input arrival process at each source node is independent The requested output port by any input packet is uniformly chosen among all output ports.

The routing within the switch is as described earlier. The output collected at a destination is C and $1 - C/I(\lambda)$ is calculated as the loss probability where $I(\lambda)$ is the total number of input packets. The simulation results are analyzed for a confidence interval of 95% and the variation in the results are within 3% of the mean value. For clarity, we do not show error bars on the graphs for clarity except in Figures 8 and 9, where we compare simulation and analytical results.

4.1 Performance Results Under Non-faulty Condition

For the non-faulty case, $\lambda = 1$ is used to simulate the performance of the network under heavy load, where each source always has a packet to send. We also vary both d and d'. Figure 2 shows the throughput obtained for different priority packets. Figure 3 depicts the results of packet loss probability of DIRSMIN and the dilated Omega networks. The loss probability decreases as the dilation degree increases for a 256×256 network. The simulations also show that the performance of d-dilated Omega network is in between that of $(d, d/2)$-DIRSMIN and (d, d)-DIRSMIN of the same size. So at equal complexity, DIRSMIN performs better than the corresponding size Omega network. We also find that the number of packets with priority 0 passed by a (d, d)-DIRSMIN is greater than that passed by a d-dilated Omega Network. Therefore even without using multiple priorities, (d, d)-DIRSMIN performs better than d-dilated Omega. It is also observed that as

the dilation degree increases, the performance difference between the two priority scheme and four priority scheme decreases. In almost all the cases, the two priority scheme performs well enough in terms of throughput.

4.2 Performance Results Under Faulty Conditions

To study the effects of faults, we assume the following fault model: (1) any SE, including those at the last stage, can fail; and (2) faulty SEs are unusable. In our simulation, we assume that no fault diagnosis is performed and the packets sent to a faulty SE are lost. The corresponding faulty Omega network is also simulated for comparison purpose. We simulate two types of fault situations. The first one assumes that one arbitrary SE at the last stage is faulty. The second one assumes that one whole subnetwork is faulty.

For the two fault situations, the packet loss probabilities of DIRSMIN and dilated Omega as a function of the dilation degree under full loads are shown in Figures 4 and 5. The simulations show that both a $(d, d/2)$-DIRSMIN and a (d, d)-DIRSMIN perform better than a d-dilated Omega in the presence of faulty SEs. In general, we find that a (d, d')-DIRSMIN, $d/2 \leq d' \leq d$, performs much better than a dilated Omega in the presence of faults. A (d, d)-DIRSMIN performs the best.

The throughput of a dilated Omega network in the presence of faults is limited by how many SEs become faulty. Each SE fault disconnects a set of sources from a set of destinations. If s is the maximum number of faulty SEs at one stage under the assumption of uniform independent traffic patterns, the minimum loss probability of a dilated Omega network is $s/(N/2)$ irrespective of the dilation degree. Thus, for a dilated 64×64 Omega network, the loss probability is limited to $1/32 = 0.031$ in the first case where one SE is faulty at the last stage, and $8/32 = 0.25$ in the second case where one fourth of the SEs at stage $(n-2)$ to 1 are faulty. In a DIRSMIN, however, no single fault can disconnect any S/D pairs. As long as the full access property (i.e., each source is able to communicate with every destination) is maintained, different performance requirements can still be met by varying the number of dilations d and d' for a given switch size.

In the second set of simulations, we use several different input traffic loads. The packet loss probabilities as a function of input traffic load for the two faulty situations are shown in Figures 6 and Figure 7, respectively. As the input load decreases, the performance improves as expected. From the simulations, we also notice that the contributions from the lower two priorities further decrease as input traffic load decreases. We conclude that the lower two priorities make significant contribution to the throughput only when dilation degrees are small (for example, $d = 4$ for a 256×256 switch) and traffic load is high. Thus, for all practical purposes, the two priority scheme works as well as the four priority scheme.

5 Performance Using Analytical Methods

The throughput of dilated banyans under the independent uniform traffic assumption was calculated analytically by Kruskal and Snir [7]. In their calculations, only one class of packets was considered. The analytical model for multiple priority classes was studied by S. Tridandapani [9] where different sources generate different class of traffic. However, in DIRSMIN, each source sends the same packet to four subnetworks with multiple priorities. Within each subnetwork, the packets are from different sources. At the last stage, priority 0 and 2 copies or priority 1 and priority 3 copies from the same source may reach the same SE. This causes traffic patterns to be correlated and the independent uniform assumption is no longer valid for priority 2 and priority 3 copies in the last stage. Furthermore, only the highest priority copy of the packet from each source is collected at each destination. The contribution of lower priority copies to the throughput is made only when there are no higher priority copies reaching the destination. Thus, the existing results can only be used within each subnetwork. A new analysis is needed to take into account the correlated traffic pattern at the last stage and destinations.

In the following, we make the same assumptions as used in our simulation. The input traffic pattern from each source is assumed to be independent and uniform. The SE operates in synchronized time slots and all packets have the same length. Each slot corresponds to the transmission time of one packet across the network.

5.1 Dilated Banyan Network

We first review the throughput and loss probability calculations for dilated banyan networks. In a banyan network, there is only one type of traffic. Let us define $R_m(j)$ to be the probability that j packets are forwarded to a tagged output of a SE at stage m, where $0 \leq j \leq d$ for a dilation degree of d. For convenience, we renumber the stages 1 to n from the input stage to the output stage in the following manner. Notice that this numbering is different than that used in the previous section.

The probability that j packets reach the input of a SE element at stage $m+1$ is

$$S_{m+1}(j) = \sum_{i=0}^{j} R_m(i) R_m(j-i)$$

. The probability that l of these j packets are destined for a tagged output is

$$\binom{j}{l} \left(\frac{1}{2}\right)^{l} \left(\frac{1}{2}\right)^{j-l} = \binom{j}{l} \left(\frac{1}{2}\right)^{j} = \binom{j}{l} 2^{-j}$$

. Thus,

$$R_{m+1}(j) = \begin{cases} \sum_{k=j}^{2d} S_{m+1}(k) \binom{k}{j} 2^{-k} & \text{, j < d} \\ \sum_{k=d}^{2d} S_{m+1}(k) \sum_{z=d}^{k} \binom{k}{z} 2^{-k} & \text{, j = d} \end{cases} \tag{1}$$

The boundary conditions at the input are given by the following.

$$R_0(j) = \begin{cases} \lambda & \text{j=1} \\ 1.0 - \lambda & \text{j=0} \\ 0 & \text{j} \neq 0, 1 \end{cases} \tag{2}$$

The packet loss probability is then given by

$$P_{loss} = 1 - \frac{\sum_{j=0}^{d} j R_n(j)}{\lambda}.$$

5.2 DIRSMIN

For analysis, we renumber the stages in (d, d')-DIRSMIN also from 1 to $n-1$ from the input stage to the output stage in this section. Recall that the dilation degree of the internal link is d and each of the four output links at the last stage is d'. The derivation for priority 0 copies can be done in exactly the same way as in the case of dilated networks. We use notations $R_m(p)(j)$ and $S_m(p)(j)$ instead of $R_m(j)$ and $S_m(j)$, respectively, to specify priority p in the following discussion.

From stage 1 to stage $n-2$, the same equations derived in the previous section can be used to calculate $R_m(0)(j)$ and $S_m(0)(j)$ for priority 0. However, the last stage (stage $n-1$) in (d, d')-DIRSMIN performs 2×4 switching. The probability that l out of j priority 0 packets are destined to a particular output at the last stage is

$$\binom{j}{l} \left(\frac{1}{4}\right)^l \left(1 - \frac{1}{4}\right)^{j-l}.$$

Thus,

$$R_{n-1}(0)(j) = \begin{cases} \sum_{k=j}^{2d} S_{n-1}(0)(k) \binom{k}{j} \left(\frac{1}{4}\right)^j \left(\frac{3}{4}\right)^{k-j} & , \mathrm{j} < d' \\ \sum_{k=d}^{2d} S_{n-1}(0)(k) \sum_{z=d}^{k} \binom{k}{z} \left(\frac{1}{4}\right)^z \left(\frac{3}{4}\right)^{k-z} & , \mathrm{j} = d' \end{cases} \tag{3}$$

Therefore, for $j = d'$

$$S_n(0)(j) = \sum_{i=0}^{j} R_{n-1}(0)(i) R_{n-1}(0)(j-i).$$

To calculate the throughput of the lower priority copies, we define $R_m(i)(b, c)$ for $(1 \leq i \leq 3)$ to be the probability that c packets of priority i and b packets of priority $i-1$ or higher are forwarded to an output address of a SE at stage m. As defined earlier, priority 0 is the highest priority and priority 3 is the lowest priority.

Since traffic to each subnetwork is identical and packets within each subnetwork are from independent sources, the independent uniform traffic assumption holds for all of the priority

copies from stage 1 to stage $n-2$. The probability that f copies of priority i and g copies of priority $i-1$ or higher reach a SE at stage $m+1$ is

$$S_{m+1}(i)(f,g) = \sum_{s=0}^{f}\sum_{t=0}^{g} R_m(i)(s,t)R_m(i)(f-s,g-t). \tag{4}$$

The probability that k of these g packets of priority i and j of these f packets of higher priorities are destined to a particular output is

$$\binom{f}{j}\left(\frac{1}{2}\right)^f \binom{g}{k}\left(\frac{1}{2}\right)^g.$$

Thus, for $j < d$ and $j+k < d$,

$$R_{m+1}(i)(j,k) = \sum_{g=k}^{2d}\sum_{f=j}^{2d-g} S_{m+1}(i)(f,g)\binom{g}{k}\left(\frac{1}{2}\right)^g \binom{f}{j}\left(\frac{1}{2}\right)^f \tag{5}$$

For $k < d$ and $j+k = d$,

$$R_{m+1}(i)(d-k,k) = \sum_{g=k}^{2d}\sum_{f=d-k}^{2d-g} S_{m+1}(i)(f,g)\binom{f}{d-k}\left(\frac{1}{2}\right)^f \sum_{z=k}^{g}\binom{g}{z}\left(\frac{1}{2}\right)^g \tag{6}$$

At the last stage, each SE is performing 2×4 routing and each output port of a SE has d' links. The probability that k of g copies of priority i and j of f copies of priority $> i$ are destined to a specific output is

$$\binom{f}{j}\left(\frac{1}{4}\right)^j\left(1-\frac{1}{4}\right)^{f-j}\binom{g}{k}\left(\frac{1}{4}\right)^k\left(1-\frac{1}{4}\right)^{g-k}.$$

Then, Equations 5 and 6, respectively, change to Equations 7 and 8 as follows.

For $j < d'$ and $j+k < d'$,

$$R_{n-1}(i)(j,k) = \sum_{g=k}^{2d}\sum_{f=j}^{2d-g} S_{n-1}(i)(f,g)\binom{f}{j}\left(\frac{1}{4}\right)^j\left(\frac{3}{4}\right)^{f-j}\binom{g}{k}\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{g-k}. \tag{7}$$

For $k < d'$ and $j+k = d'$,

$$\begin{aligned} R_{n-1}(i)(d'-k,k) &= \sum_{g=k}^{2d}\sum_{f=d'-k}^{2d-g} S_{n-1}(i)(f,g)\binom{f}{d'-k}\left(\frac{1}{4}\right)^{d'-k}\left(\frac{3}{4}\right)^{f-(d'-k)} \\ &\quad \cdot\sum_{z=k}^{g}\binom{g}{z}\left(\frac{1}{4}\right)^z\left(\frac{3}{4}\right)^{g-z}. \end{aligned} \tag{8}$$

The throughput for priority 0 packets is given by

$$T(0) = 2\sum_{j=0}^{d'} jR_{n-1}(0)(j).$$

The factor of 2 accounts for the fact that there are two groups of d' links from two different SEs at the last stage to each destination. The probability that the priority 0 copy of a packet from an input node passes through the network is simply $P_{n-1}(0) = T(0)/\lambda$.

The boundary conditions for priority $i = 1,\ 2$ and 3 copies at the input are

$$R_0(i)(j,k) = \begin{cases} \binom{i}{j}(1-\lambda)^{i-j}\lambda^{j+1} & , 0 \leq j \leq i, \text{ k=1} \\ \binom{i}{j}(1-\lambda)^{i-j+1}\lambda^{j} & , 0 \leq j \leq i, \text{ k=0} \\ 0 & , \text{otherwise} \end{cases}$$

To calculate the throughput of priority 1 copies, we observe that a priority 1 copy from one particular source is accepted by its destination only when it successfully reaches its destination and the corresponding priority 0 copy does not make it to the destination because of contentions. We first calculate the probability that k $(0 \leq k \leq d')$ priority 1 copies reach one particular output of a SE at the last stage as,

$$R'_{n-1}(1)(k) = \sum_{j=0}^{d'-k} R_{n-1}(1)(j,k).$$

Then,

$$P_{n-1}(1) = 2\sum_{k=0}^{d'} kR'_{n-1}(1)(k)/\lambda$$

gives the probability that a priority 1 copy passes through the network. The probability that a priority 1 copy of an input packet is the only copy at the destination is $P_{n-1}(1)(1 - P_{n-1}(0))$. Notice that $(1 - P_{n-1}(0))$ is the probability that the corresponding priority 0 copy does not reach the destination due to contentions. Even though the four subnetworks are isomorphic, the copies from each source are assigned with different priorities inside different subnetworks, forming different traffic patterns inside different subnetworks. Suppose both priority 1 and

priority 0 copies reach the destination with no loss, then the probability that priority 1 is the only copy at the destination is 0, which is correctly described by the expression above. The net throughput of priority 1 copies can be calculated as

$$T(1) = 2\sum_{k=1}^{d'}\sum_{y=1}^{k}\binom{k}{y} y(P_{n-1}(1)(1-P_{n-1}(0)))^y(1-P_{n-1}(1))^{k-y}R'_{n-1}(1)(k) \qquad (9)$$

The calculations of the throughput of priority 2 and priority 3 copies are further complicated by the fact that priority 2 (0) and priority 3 (1) copies of the same input packet, if both are successful, reach the same SE at the last stage. Thus, the independent uniform traffic assumption cannot be used directly for priority 2 and 3 traffic at the input of the last stage. We first calculate the number of distinct priority 2 (3) copies that do not have their corresponding higher priority copies reaching the same SE at last stage. This can be done in the same way as above in the calculation of distinct priority 1 copies. We use the independent uniform traffic assumption on these distinct priority 2 (3) copies to calculate the net throughput of priority 2 (3).

The probability that a priority i ($i = 2, 3$) copy passes through the network up to the input of the last stage is

$$P_{n-2}(i) = \sum_{k=0}^{d}\sum_{j=0}^{d-k} kR_{n-2}(i)(j,k)/\lambda.$$

The probability that k distinct copies of priority 2 and j copies of higher priorities reach an input port of a SE at the last stage is

$$\overline{R}_{n-2}(2)(j,k) = \sum_{y=k}^{d} R_{n-2}(2)(j,y)\binom{y}{k}[P_{n-2}(2)(1-P_{n-2}(0))]^k(1-P_{n-2}(2))^{y-k} \qquad (10)$$

The probability that k distinct copies of priority 3 and j copies of higher priorities reach an input port of a SE at the last stage is given simply by substituting quantities for priority 2 with those for priority 3, and quantities for priority 0 with those for priority 1 in the above equation.

The independent uniform traffic assumption can now be applied on these distinct priority 2 and 3 copies at the input of the last stage. Equation 4 can be used first to calculate from $\overline{R}_{n-2}(2)(j,k)$ the probability $\overline{S}_{n-1}(i)(j,k)$ ($i = 2, 3$) that k distinct priority i copies and j higher

priority copies reach the input of a SE at the last stage. From that $\overline{R}_{n-1}(i)(j,k)$ $(i = 2,3)$, which is the probability that k distinct priority i packets and j packets of higher priorities are forwarded to an output port of a SE at the last stage, can be calculated using Equation 7 and Equation 8. The total number of priority 2 copies reaching the destination, which are distinct from their corresponding priority 0 copies, is

$$\overline{T}(2) = 2\sum_{k=0}^{d'}\sum_{j=0}^{d'-k} kR_n(2)(j,k).$$

Similarly the total number of priority 3 copies reaching a destination, which are distinct from their corresponding priority 1 copies, is

$$\overline{T}(3) = 2\sum_{k=0}^{d'}\sum_{j=0}^{d'-k} kR_n(1)(j,k)$$

At the destination, some of these priority 2 (3) copies again are dropped because the corresponding priority 1 (0) copy of the same packet from the conjugate SE reaches the destination. The probability that this happens to priority 2 copies can be estimated by

$$T(1)/(\lambda - T(0)),$$

where $\lambda - T(0)$ is the remaining traffic that is not passed by priority 0 copies and $T(1)$ is the traffic passed by priority 1 copies among these remaining traffic. Finally, the net throughput of priority 2 copies is given by

$$T(2) = \overline{T}(2)(1 - T(1))/(\lambda - T(0))).$$

We find that the net throughput of priority 2 copies is always a few orders of magnitude smaller than that of priority 0 copies for DIRSMIN, and therefore, it can be neglected in estimating the net throughput of priority 3 copies. Notice that out of $\lambda P_{n-1}(1)$ priority 1 copies that get to a destination, $T(1)$ are distinct from priority 0 copies. From this, the distinct priority 3 copies which are accepted by a destination can be estimated by

$$T(3) = \overline{T}(3)\frac{T(1)}{\lambda P_{n-1}(1)}.$$

Finally the throughput of the network is

$$T = T(0) + T(1) + T(2) + T(3).$$

The packet loss probability of DIRSMIN is

$$P_{loss} = 1 - \frac{T}{\lambda}.$$

5.3 Numerical Results

The results obtained above are checked against the simulation results and are found to be consistent. One comparison of results for a 64×64 network is shown in Figure 8 and 9. The performance of a (d, d')-DIRSMIN and a dilated Omega network at full load ($\lambda = 1$) and variable loads is shown in Figures 10 and 11 for a 256×256 size network. We observe that the performance of a d-dilated Omega lies in between a $(d, d/2)$-DIRSMIN and a (d, d)-DIRSMIN. The performance of DIRSMIN improves as both d and d' increase and the role of the two low priorities diminishes.

6 Reliability Analysis of DIRSMIN

The all-terminal reliability, $R(t)$, is one of the most important measures of the effectiveness of a fault-tolerant scheme employing redundancy. This is the probability that there exists a path between each source and every destination. SE failures are random and independent events. Exact analysis of reliability in general is known to be NP-hard [18]. Normally only analytical bounds on reliability can be obtained. Monte Carlo simulations have been used to get more accurate numerical results.

Due to the unique path property of Omega type banyan networks, the all-terminal reliability for such networks is $r^{N/2 \log N}$, where $N/2 \log N$ is the total number of SEs in the banyan network. The reliability diminishes very quickly with the increasing size of the network. The multiple paths of DIRSMIN tremendously help improve the reliability. We assume that all the SEs in the switch have the same reliability $r(t)$.

6.1 Network Reliability of DIRSMIN

We note from Figure 1g that DIRSMIN can be redrawn as four SE-disjoint sub-banyans linked at the last stage. The whole network consists of two disjoint subsystems, each consisting of two sub-banyans linked by N/4 SEs at the last stage. In Figure 1h, one of the two subsystems consists of subnetworks 0 and 2 and the $N/4$ SEs at the last stage that they are connected to, and the other one consists of subnetworks 1 and 3 and their corresponding SEs at the last stage. The reliability of two identical subsystems in parallel is

$$R = 1 - (1 - R_{sub})^2, \tag{11}$$

where R_{sub} is the reliability of each subsystem. In the following, we first estimate the bounds for one subsystem and then use the above equation to get the bounds for DIRSMIN.

To estimate the lower reliability bound of one subsystem, we observe that the full access property of a subsystem is maintained as long as (1) all the $N/4$ SEs in the last stages are not faulty, and (2) at least one of the two complete subnetworks is fault-free. Thus, a conservative lower bound on the all-terminal reliability of one subsystem is given by

$$R_{sub} > r^{N/4} \times (1 - (1 - r^{N'})^2). \tag{12}$$

In the above equation, $N' = N/8(\log N - 2)$ is the total number of SEs in each of the four subnetworks of DIRSMIN.

To obtain the upper bound for a subsystem, we first observe that each SE in a particular stage from stage 1 to stage $(n-2)$ has one conjugate SE within one subsystem. Two SEs are conjugate if they occupy corresponding positions in the two subnets to which they belong. The subsystem fails if a conjugate pair of SEs fails. The subsystem is operational as long as no conjugate pair of SEs fails and no SE in the last stage fails. In this estimation, since there are many combinations of failed SEs which cause the subsystem to fail other than conjugate SE pair, the system reliability is overestimated. Therefore, the upper bound for the all-terminal reliability of a subsystem is

$$R_{sub} < r^{N/4} \times (1 - (1 - r)^2)^{N'}. \tag{13}$$

Finally, from Equations 11 and 12, we get the lower bound of the all-terminal reliability of DIRSMIN as

$$R > 1 - \left(1 - r^{N/4} \times (1 - (1 - r^{N'})^2)\right)^2.$$

From Equations 11 and 13, we get the higher bound of the all-terminal reliability of DIRSMIN as

$$R < 1 - \left(1 - r^{N/4} \times (1 - (1 - r)^2)^{N'}\right)^2.$$

6.2 Numerical Results and Comparison with SEN+

Using extra stages to create redundant paths between any S/D pairs has been proposed in the literature. The most robust and reliable of these networks, SEN+, has been analyzed in [19]. We, therefore, compare the all-terminal reliability of SEN+ with DIRSMIN. Analytical bounds on reliability for SEN+ have been estimated in [19]. In SEN+, two SE-disjoint subnetworks exist between the input and output stages. No SE faults at the first and last stage can be tolerated. The upper and lower all-terminal reliability bounds were obtained as

$$R < r^N \times [1 - (1 - r)^2]^{N'}$$

and

$$R > r^N \times (1 - (1 - r^{N'})^2),$$

respectively, where $N' = (N/4)(\log N - 1)$ is the number of SEs in each of the two subnets.

Figures 12 and 13 depict the comparisons of the all-terminal reliability of dilated ELMIN and SEN+ using the above relations. Figure 12 shows the dependence of network reliability R on SE reliability for a 64×64 network. Figure 13 shows the network reliability for different size networks using a fixed reliability SE with $r = 0.999$. In all cases, the lower reliability bounds of DIRSMIN are considerably higher than the higher reliability bounds of SEN+. Thus, the reliability of DIRSMIN is much higher than that of SEN+. The main reason for this is that there are four SE-disjoint subnetworks in DIRSMIN whereas there are only two in SEN+. DIRSMIN can tolerate faults at any stage including the last stage which is not the case for SEN+.

7 Conclusions

We have developed a fault tolerant fast packet switch design, the (d, d')-DIRSMIN, which uses dilation to improve performance and fault tolerance of a network. This new network is capable of providing low packet loss probability and high reliability with very little hardware overhead compared to d-dilated banyan networks. Under non-faulty conditions, both simulation and analytical results show that a (d, d)-DIRSMIN performs better than the original dilated banyan network with the same SE complexity. Under faulty conditions, simulation results show that a (d, d')-DIRSMIN performs much better than a d-dilated Omega network. In these cases, a (d, d')-DIRSMIN yields monotonically decreasing loss probability as a function of the dilation degree, whereas a d-dilated banyan network cannot provide connection between certain S/D pairs and the loss probability is bounded depending on how many faults are present in the network.

A multiple priority scheme allows us to explore alternate paths simultaneously which results in higher throughput and reliability under both fault-free and faulty conditions. A (d, d')-DIRSMIN tolerates multiple SE faults inside the network, including SEs at the input and output stages. It is shown that the reliability of DIRSMIN is considerably higher than that of SEN+.

References

[1] C. Wu and T. Feng, "On a Class of Multistage Interconnection Networks," *IEEE Trans. Comput.*, C-28, pp. 694-702, Aug. 1980.

[2] Y. N. Hui, and E. Arthurs, "A Broadband Packet Switch for Integrated Transport," *IEEE Journal on Selected Areas in Communications*, Vol. 5, No. 8, pp. 1264-1273, Oct 1987.

[3] X. Jiang and J. S. Meditch, "Integrated services fast packet switching," in *Proc. IEEE GLOBECOM '89*, Nov. 1989, pp. 1478-1482.

[4] F. A. Tobagi and T. Kwok, "Architecture, Performance, and Implementation of Tandem Banyan Fast Packet Switch," *IEEE J. on Select. Areas Commun.*, Vol. 9, No. 8, pp. 1173-1193, Oct 1991.

[5] Y. S. Yeh, M. G. Hluchyj and A. S. Acampora, "The knockout switch: A simple modular architecture for high-performance packet switching," *IEEE J. Select. Areas Commun.*, vol. SAC-5, Oct. 1987, pp. 1274-1281.

[6] A. Huang and S. Knauer, "Starlite: A wideband digital switch," in *Proc. IEEE GLOBECOM '84*, Dec. 1984, pp. 121-125.

[7] C. P. Kruskal and M. Snir, "The Performance of Multistage Interconnection Networks for Multiprocessors," *IEEE Trans. Comput.*, C-32, pp. 1091-1098, Dec 1983.

[8] M. Kumar and J. R. Jump,"Performance of Unbuffered Shuffle-Exchange Networks", *IEEE Tran. Comput.*, vol C-35, No. 6, pp. 573-578, June 1986.

[9] S. Tridandapani and J. S. Meditch, "Priority Performance of Banyan-based Broadband-ISDN Switches," *J. of High Speed Networks*, vol.3, no.3, pp. 233-60, 1994.

[10] M. G. Hluchyj and M. J. Karol, "Queueing in High-Performance Packet Switching," *IEEE J. on Select. Areas Commun.*, vol. 6, pp. 1587-1597, Dec. 1988.

[11] G. B. Adams and H. J. Siegel, "The extra stage cube: A fault tolerant network for super-systems," *IEEE Trans. Comput.*, vol. C-31, May 1982, pp. 443-457.

[12] V. P. Kumar and S. M. Reddy, "Augmented shuffle-exchange multistage interconnection networks," *IEEE Computer*, vol. 20, June 1987, pp. 30-40.

[13] K. Padmanabhan and D. H. Lawrie, "A class of redundant path multistage interconnection network," *IEEE Trans. on Comput.*, Dec. 1983, pp. 1145-1155.

[14] C. S. Raghavendra and A. Varma, "INDRA: A class of interconnection network with redundant paths," *1984 Real Time Syst. Symp.*, Computer Society Press, Silver Spring, Md., 1984, pp. 153-164.

[15] S. B. Choi and A. K. Somani, "Design and Performance Analysis of Load-distributing Fault-tolerant Network," to appear in *IEEE Trans. on Comput.*.

[16] F. T. Leighton and B. M. Maggs, "Fast Algorithm for Routing Around Faults in Multibutterflies and Randomly Wired Splitter Networks," *IEEE Trans. on Comput.*, vol. 41, No. 5, May 1992, pp. 578-587.

[17] T. M. Zhang, "Achieving Almost Free Fault Tolerance in Dilated Banyan Network," *M.S. Thesis*, Depart. of Elect. Eng., FT-10, University of Washington, Seattle, WA 98195, 1992.

[18] M. O. Ball, "Computational Complexity of Network Reliability Analysis: An Overview," *IEEE Trans. Reliability*, R-35, pp. 230-239, Aug, 1986.

[19] J. T. Blake and K. S. Trivedi, "Multistage Interconnection Network Reliability," *IEEE Trans. Comput.*, C-38, pp. 1600-1604, Nov. 1989.

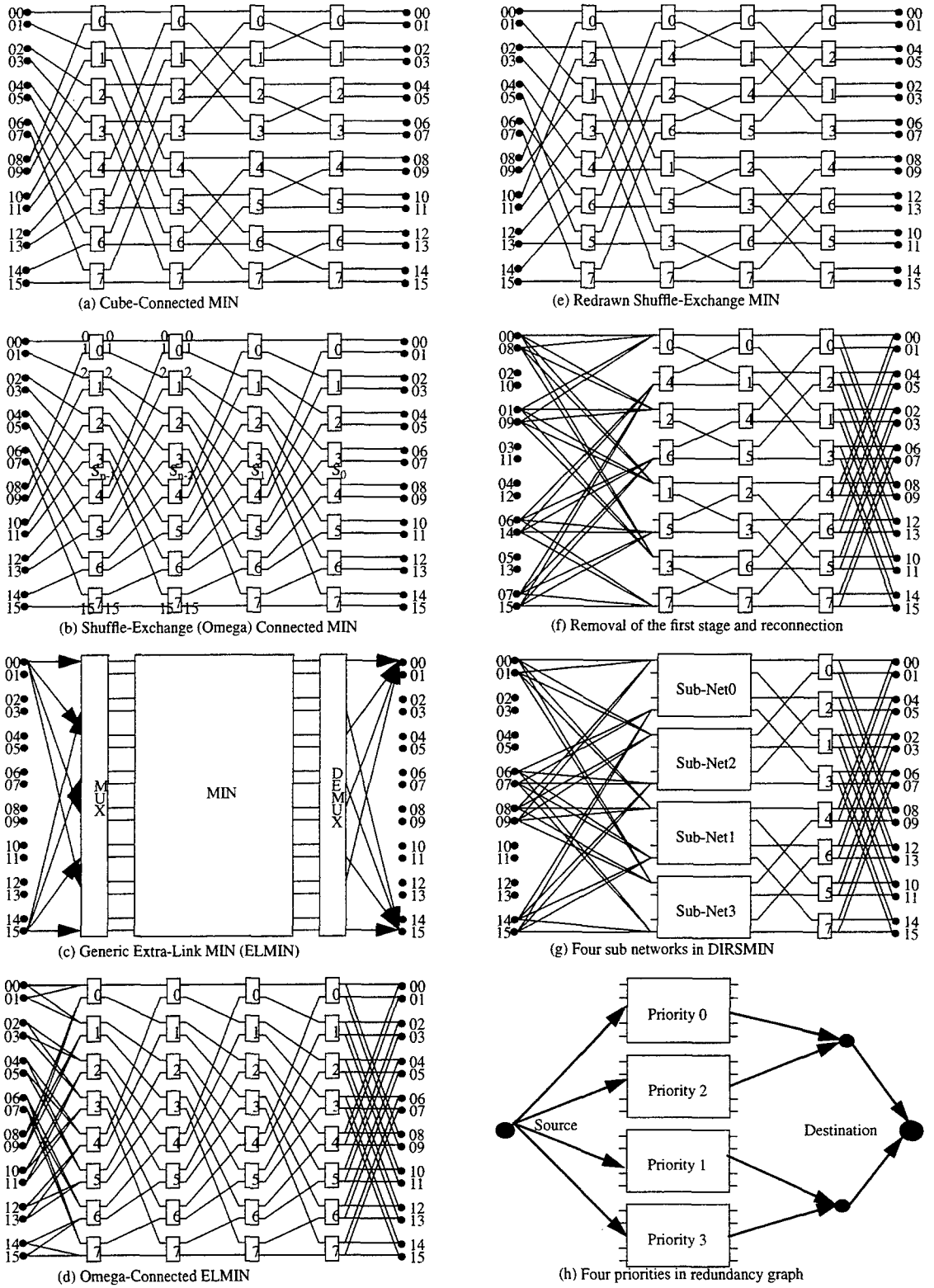

Figure 1: Various Multistage Interconnection Configurations

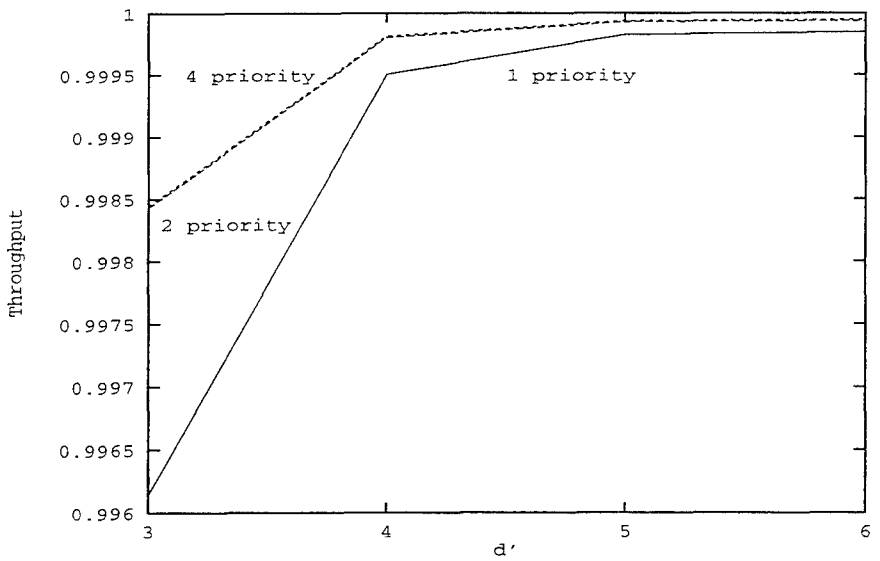

Figure 2: Simulation results for 256X256 (6,d')-DIRSMIN and Omega Networks: Throughput for different priority schemes (curves for 2 and 4 overlap).

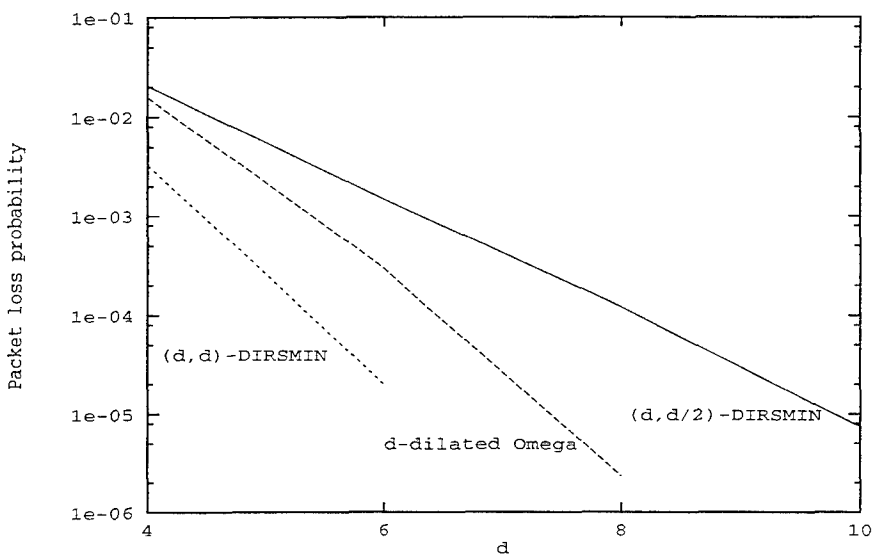

Figure 3: Simulation results for 256X256 (6,d')-DIRSMIN and Omega Networks: Packet loss probability at full load.

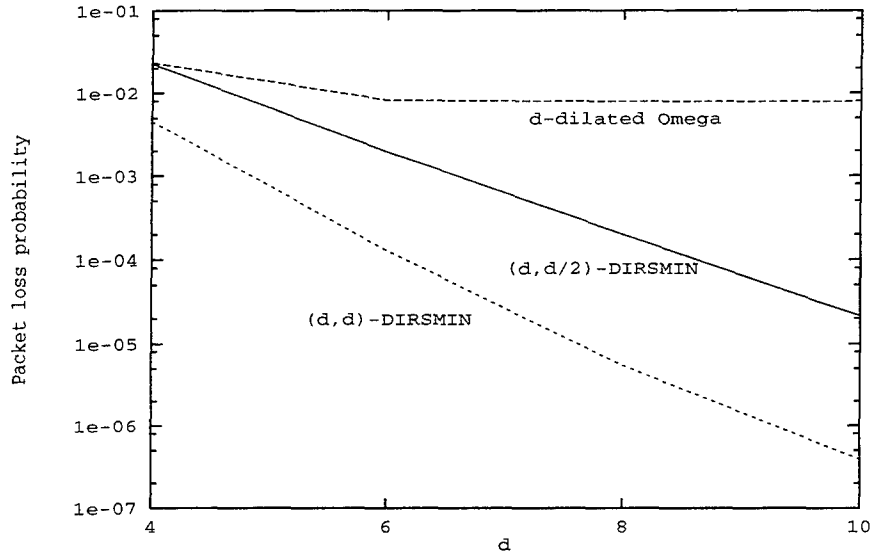

Figure 4: Packet loss probability with one faulty switch. Size 256X256. $\lambda = 1$.

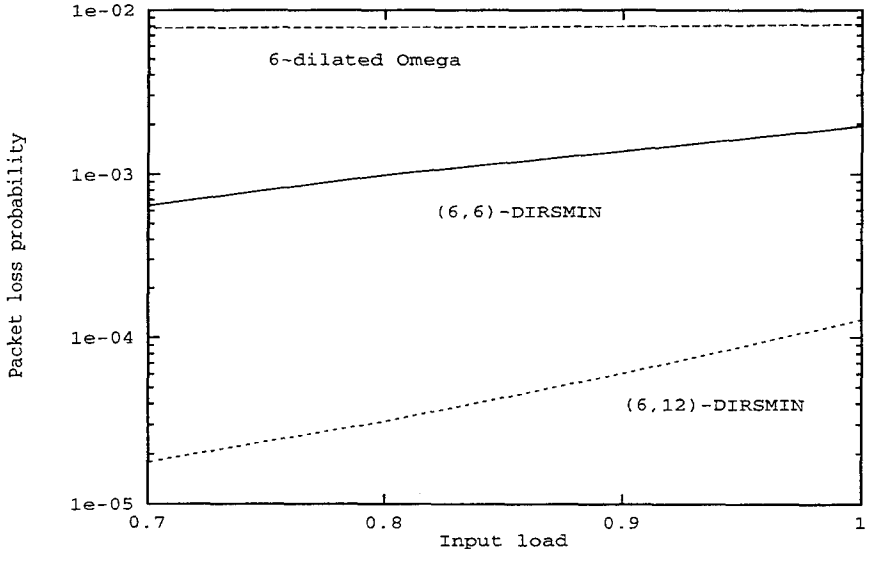

Figure 5: Packet loss probability with one faulty switch. Size 256X256. λ = variable.

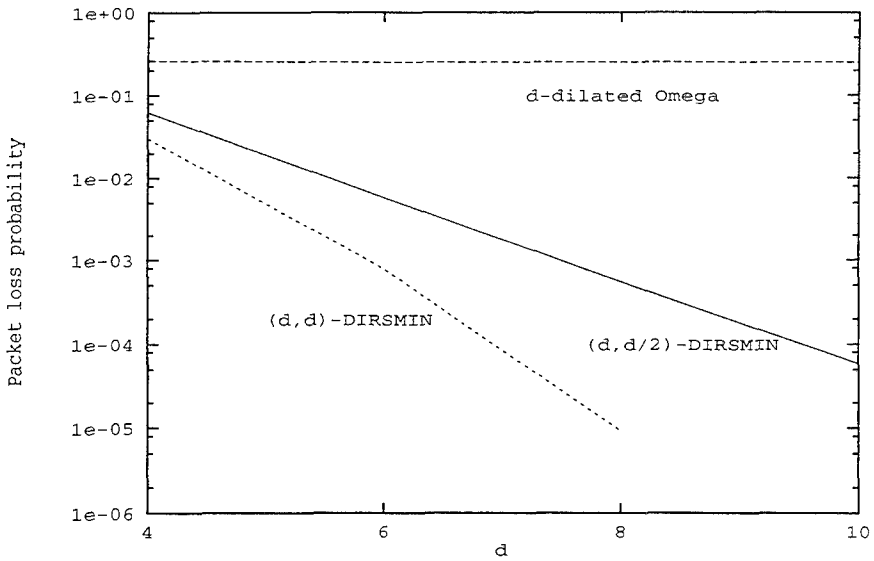

Figure 6: Packet loss probability with one faulty subnetwork. Size 256X256. $\lambda = 1$.

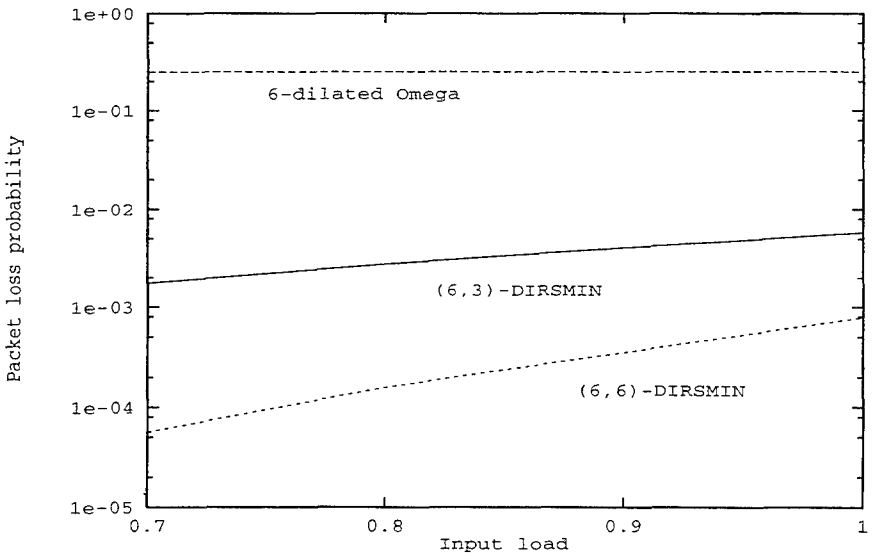

Figure 7: Packet loss probability with one faulty subnetwork. Size 256X256. λ = variable.

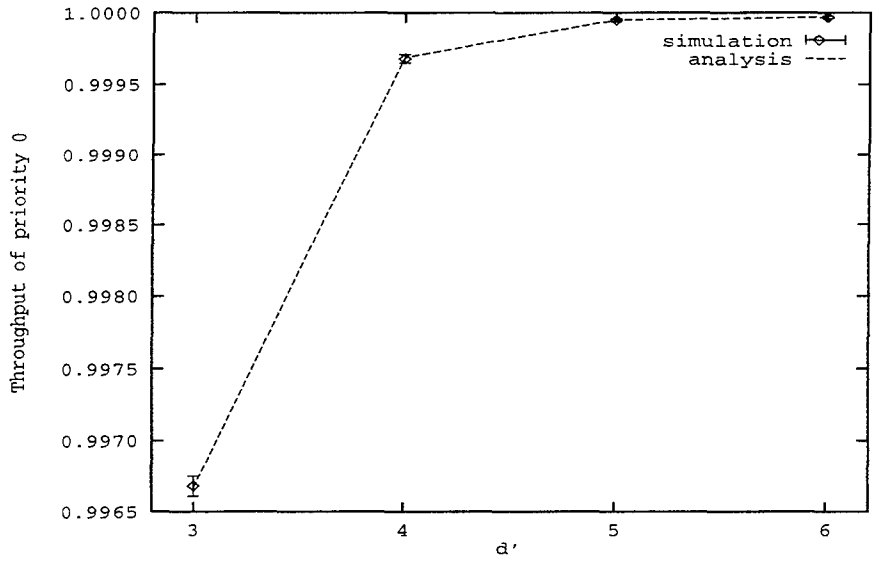

Figure 8: Comparison of Simulation and Analytical Results for 64×64 $(6, d')$-DIRSMIN. Priority 0.

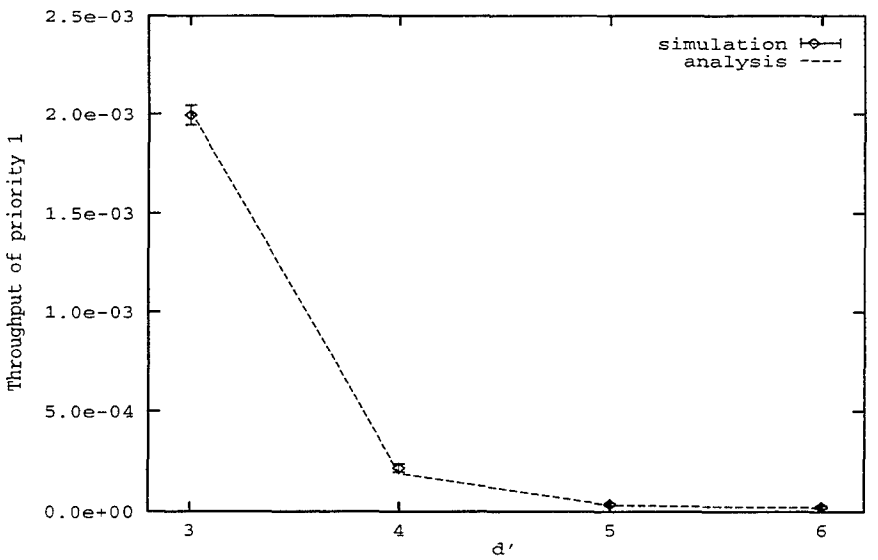

Figure 9: Comparison of Simulation and Analytical Results for 64×64 $(6, d')$-DIRSMIN. Priority 1.

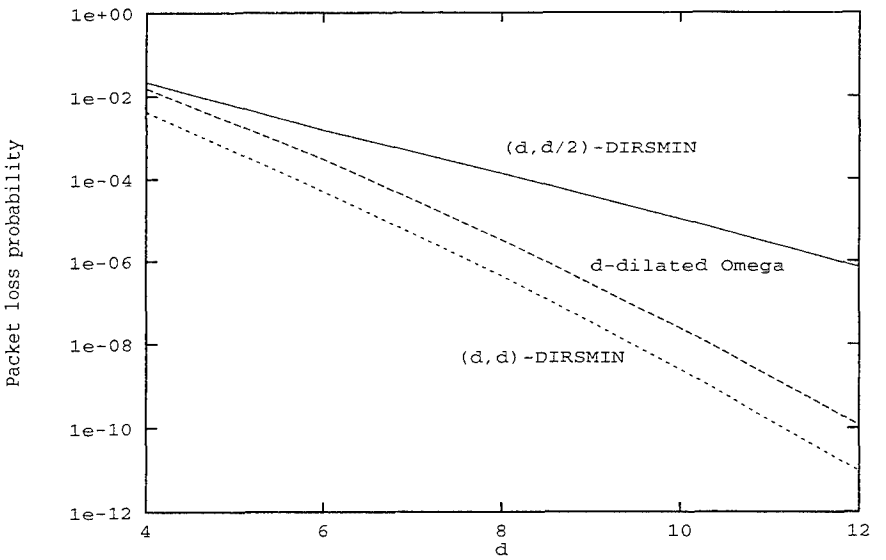

Figure 10: Analytical results of a 256×256 DIRSMIN and a dilated Omega. $\lambda = 1$.

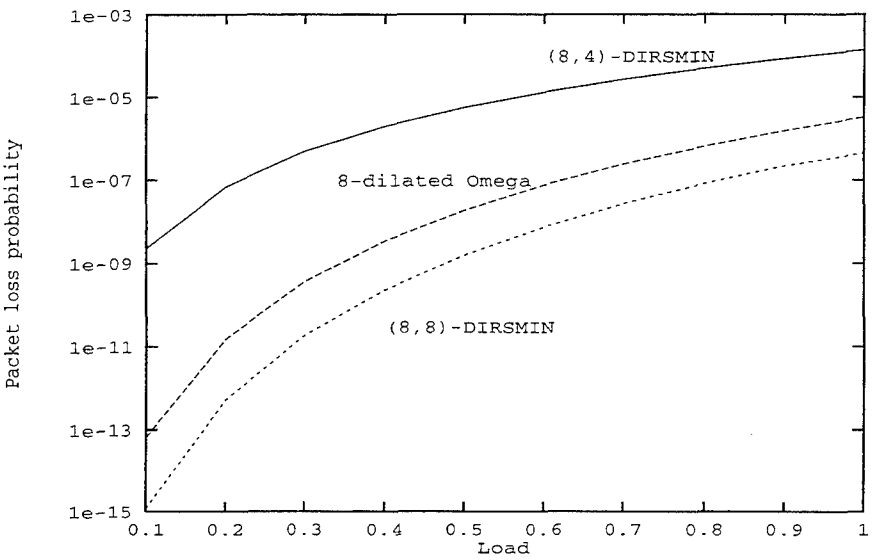

Figure 11: Analytical results of a 256×256 DIRSMIN and a dilated Omega. λ = variable.

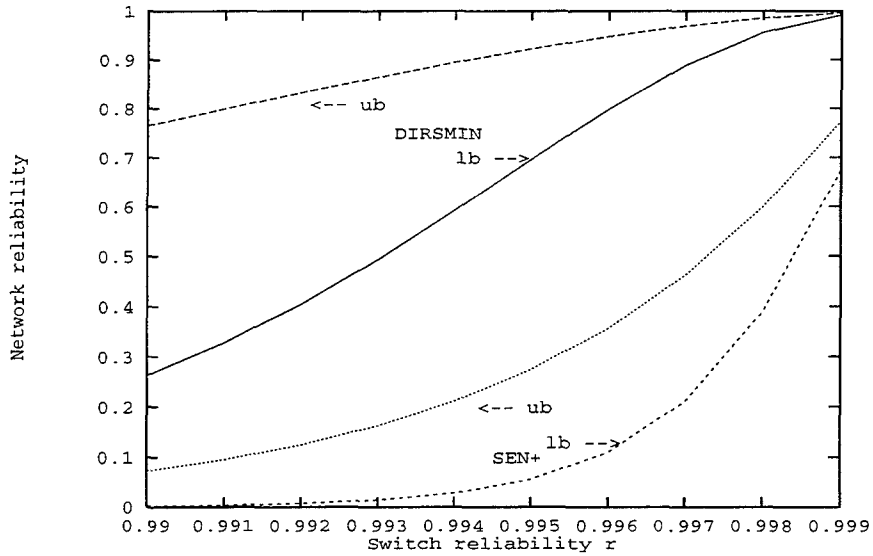

Figure 12: All-terminal reliability bounds for DIRSMIN and SEN+. $N = 64$ and r = variable

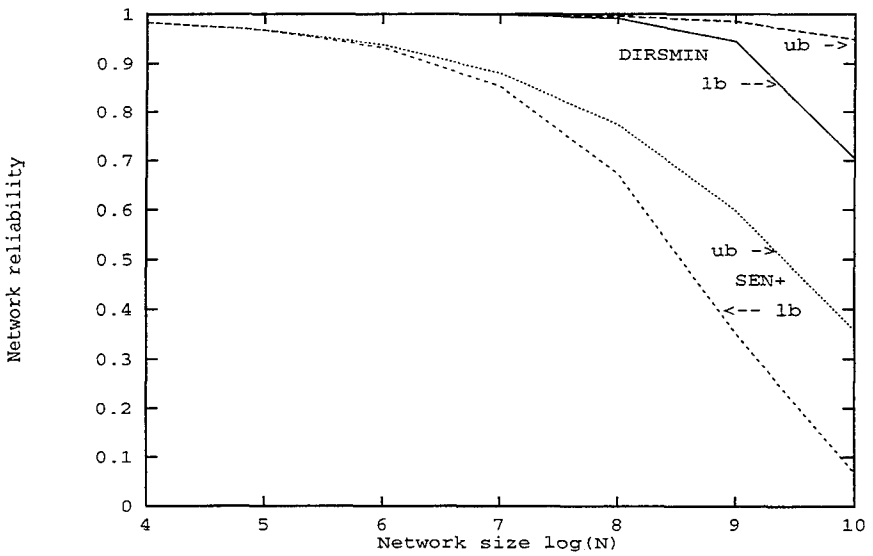

Figure 13: All-terminal reliability bounds for DIRSMIN and SEN+. $r = 0.999$.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 1995	Contractor Report

4. TITLE AND SUBTITLE
ARCHITECTURE AND PERFORMANCE ANALYSIS OF DIRSMIN: A FAULT-TOLERANT SWITCH USING DILATED REDUCED-STAGE MIN

5. FUNDING NUMBERS
C NAS1-19480
WU 505-90-52-01

6. AUTHOR(S)
Arun K. Somani
Tianming Zhang

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Computer Applications in Science and Engineering
Mail Stop 132C, NASA Langley Research Center
Hampton, VA 23681-0001

8. PERFORMING ORGANIZATION REPORT NUMBER
ICASE Report No. 95-78

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Langley Research Center
Hampton, VA 23681-0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NASA CR-198238
ICASE Report No. 95-78

11. SUPPLEMENTARY NOTES
Langley Technical Monitor: Dennis M. Bushnell
Final Report
To be submitted to IEEE Transactions on Communications

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified–Unlimited

Subject Category 60, 61

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

We develop and analyze a dilated high performance fault tolerant fast packet multistage interconnection network (MIN) in this paper. In this new design, the links at the input and the output stages of a dilated banyan-based MIN are rearranged to create multiple routes for each source-destination pair in the network after removing one stage in the network. These multiple paths are link- and node-disjoint. Fault tolerance at low latency is achieved by sending multiple copies of each input packet simultaneously using different routes and different priorities. This guarantees that high throughput is maintained even in the presence of faults. Throughput is analyzed using simulation and analysis and we show that the new design has considerably higher performance in the presence of a faulty switching element (SE) or link in comparison to dilated networks. We also analyze the reliability and show that the new design has superior reliability in comparison to competing proposals.

14. SUBJECT TERMS
Multi-Stage Interconnection Network; Dilated MIN; Extra-Link MIN (ELMIN); Dilated Reduced-Stage MIN (DIRSMIN); Low-Latency Tolerance; Reliability in Switching

15. NUMBER OF PAGES
29

16. PRICE CODE
A03

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified		